Прежде всего сформилируем условия задачи:

Анализ мутаций коронавируса из людей разных стран

Задача состоит в том, чтобы взять 9 геномов коронавируса SARS-CoV-2 из людей разных стран и 10-ой геном коронавируса SARS-CoV-1, выровнять последовательности и построить деревья тремя разными методами – расстояний, максимальной бережливости, максимального правдоподобия, без бутстрэпа (так как он не имеет смысла при таком малом количестве мутаций). Сделать вывод о том, кого вирус заразил раньше, а кого самого последнего.

Взять 2 генома самого первого после SARS-CoV-1 по дереву и самого последнего и выписать координаты 5 любых мутаций. Посмотреть, куда они попадают – в ген (какой?) или межгенное пространство (какое, между какими генами?).

Разметка генома
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/009/858/895/GCF_009858895.2_ASM985889v3/GCF_009858895.2_ASM985889v3_genomic.gff.gz

Полные геномы SARS-CoV-2 лежат тут: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049

При скачивании обращайте внимание на размер файла, должен быть ~30 кВ. Можно включить фильтр nucleotide completeness -> complete

Отчет должен содержать:

1. Скриншоты дерева.
2. Скриншоты мутаций на выравнивании, где видны координаты.
3. Список координат и места попадания.

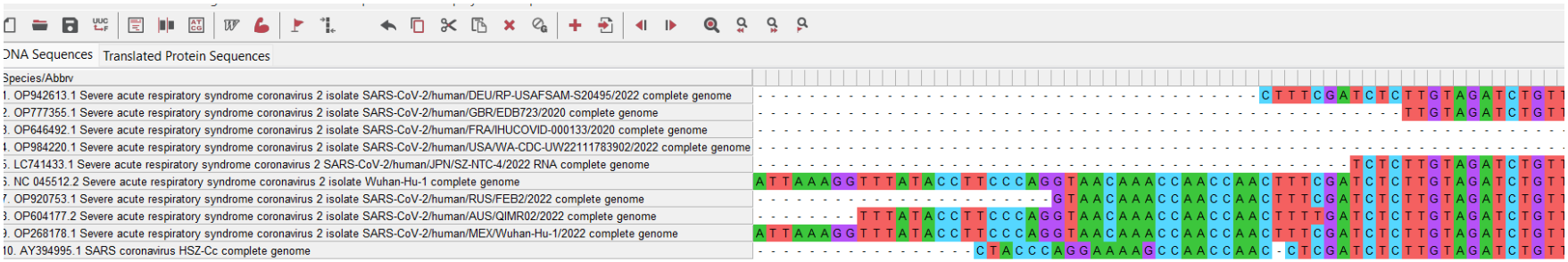
www.ncbi.nlm.nih.gov (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/)

Прежде всего выпишем те геномы, которые взяли:

1. China - NC_045512 +
2. Japan - LC741433 +
3. USA - OP984220 +
4. France - OP646492 +
5. UK - OP777355 +
6. Germany - OP942613 +
7. Russia - OP920753 +

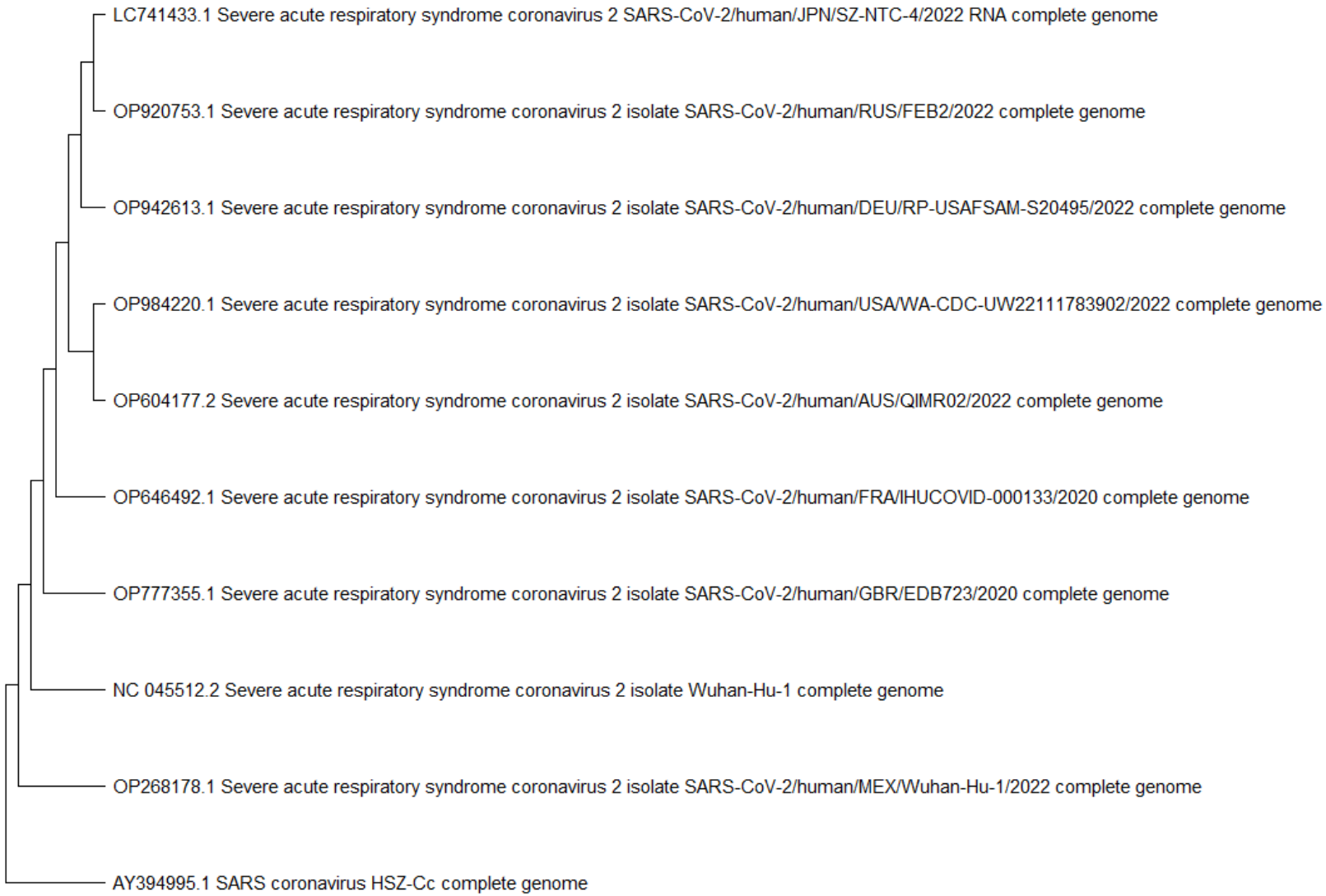
8. Mexico - OP268178 +
9. Australia - OP604177 +
10. SARS-Cov-1 (с семинара) +

Получаем выравненную последовательность:

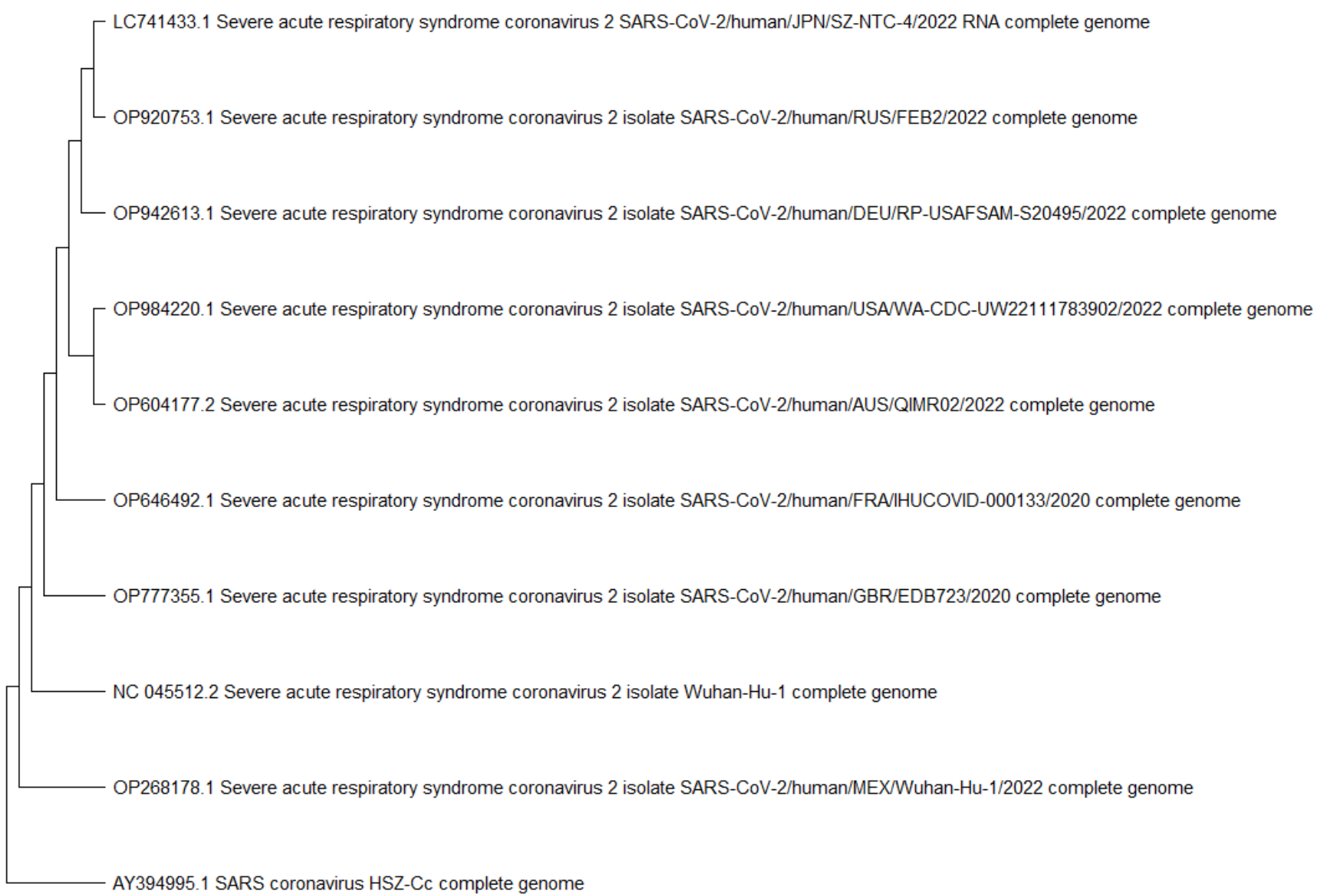


И готовые деревья:

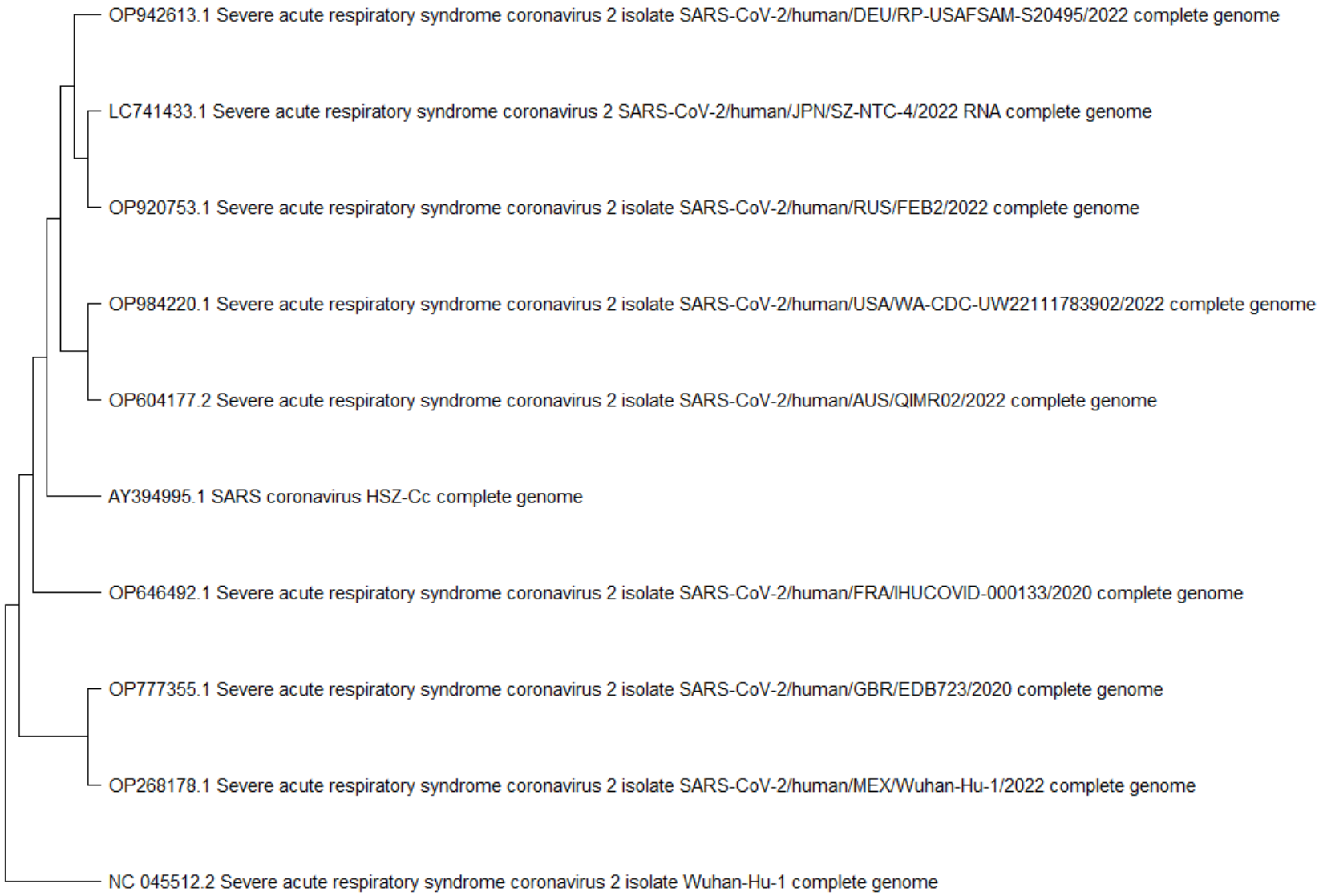
ML:


LC741433.1 Severe acute respiratory syndrome coronavirus 2 SARS-CoV-2/human/JPN/SZ-NTC-4/2022 RNA complete genome
OP920753.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/RUS/FEB2/2022 complete genome
OP942613.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/DEU/RP-USAFSAM-S20495/2022 complete genome
OP984220.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/WA-CDC-UW22111783902/2022 complete genome
OP604177.2 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/AUS/QIMR02/2022 complete genome
OP646492.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/FRA/IHUCOVID-000133/2020 complete genome
OP777355.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/GBR/EDB723/2020 complete genome
NC 045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 complete genome
OP268178.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/MEX/Wuhan-Hu-1/2022 complete genome
AY394995.1 SARS coronavirus HSZ-Cc complete genome

NJ:


LC741433.1 Severe acute respiratory syndrome coronavirus 2 SARS-CoV-2/human/JPN/SZ-NTC-4/2022 RNA complete genome
OP920753.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/RUS/FEB2/2022 complete genome
OP942613.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/DEU/RP-USAFSAM-S20495/2022 complete genome
OP984220.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/WA-CDC-UW22111783902/2022 complete genome
OP604177.2 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/AUS/QIMR02/2022 complete genome
OP646492.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/FRA/IHUCOVID-000133/2020 complete genome
OP777355.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/GBR/EDB723/2020 complete genome
NC 045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 complete genome
OP268178.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/MEX/Wuhan-Hu-1/2022 complete genome
AY394995.1 SARS coronavirus HSZ-Cc complete genome

MP:



Вывод:

В целом, для методов ML и NJ картина схожая - дальше всех друг от друга находится sars-cov-1 (Китайский, стартовый вариант ковида) и вариант из Японии, Германии. Для метода максимальной бережливости картина схожая, но разница есть - в нижней ветке хоть и стоит китайский вариант генома, но для варианта sars-cov-2, что немного странно.

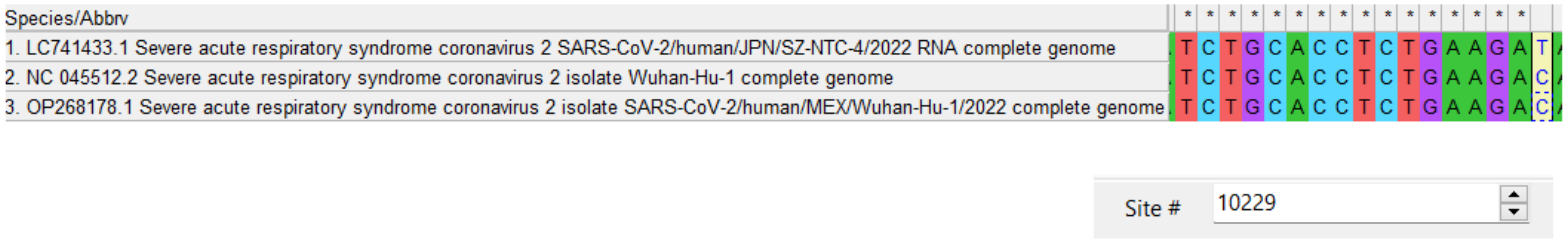
Смотря в целом, можно отметить, что соседство в филогенетическом дереве отвечает соседству территориальному (за некоторыми исключениями, но тут скорее вопрос в какое время взят геном и у кого)

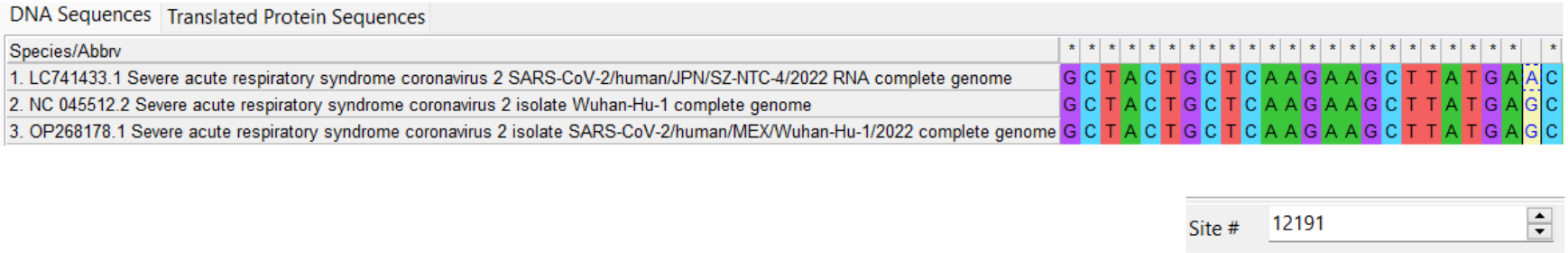
Мутации:

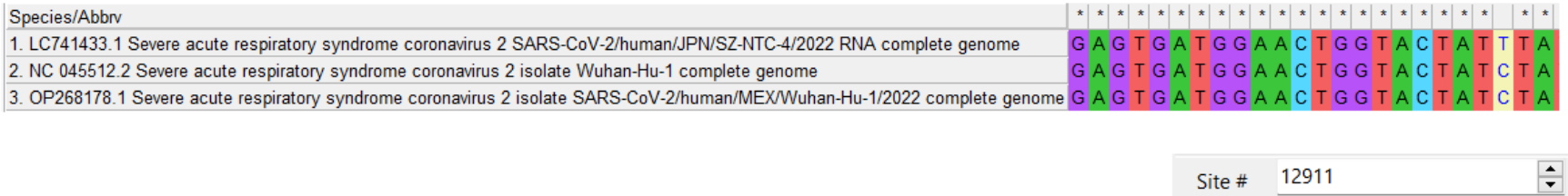
Теперь возьмем 2 генома самого первого после SARS-CoV-1 по дереву и самого последнего и выпишем координаты 5 любых мутаций. Соотвественно нас интересуют:

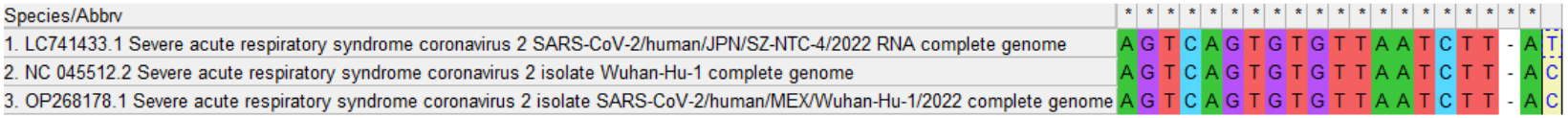
1. OP268178
2. NC_045512
3. LC741433

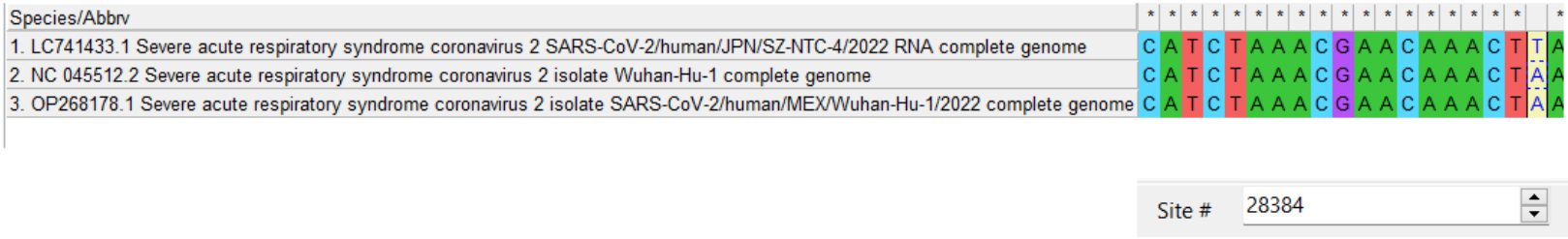
Результаты поиска:











Вывод:

Соотвественно мы можем видеть, что первые три мутации произошли в гене "**ORF1ab**", после чего я смог найти, например, мутацию в гене с названием "**S**" и еще мутацию, например в гене "N".

Пользовался следующей разметкой:

GCF_009858895.2_ASM985889v3_genomic.gff

```
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=2697049
NC_045512.2 RefSeq  region  1   29903   .   +   .   ID=NC_045512.2:1..29903;Dbxref=taxon:2697049;collection-date=Dec-2019;country=China;gb-acronym=SARS-CoV-2;gbkey=Src;genome=genomic;isola
NC_045512.2 RefSeq  five_prime_UTR  1   265 .   +   .   ID=id-NC_045512.2:1..265;gbkey=5'UTR
NC_045512.2 RefSeq  gene    266 21555   .   +   .   ID=gene-GU280_gp01;Dbxref=GeneID:43740578;Name=ORF1ab;gbkey=Gene;gene=ORF1ab;gene_biotype=protein_coding;locus_tag=GU280_gp01
NC_045512.2 RefSeq  CDS 266 13468   .   +   0   ID=cds-YP_009724389.1;Parent=gene-GU280_gp01;Dbxref=Genbank:YP_009724389.1,GeneID:43740578;Name=YP_009724389.1;Note=pplab%3B translated by -
NC_045512.2 RefSeq  CDS 13468   21555   .   +   0   ID=cds-YP_009724389.1;Parent=gene-GU280_gp01;Dbxref=Genbank:YP_009724389.1,GeneID:43740578;Name=YP_009724389.1;Note=pplab%3B translated
NC_045512.2 RefSeq  mature_protein_region_of_CDS    266 805 .   +   .   ID=id-YP_009724389.1:1..180;Note=nsp1%3B produced by both ppla and pplab;Parent=cds-YP_009724389.1;gbkey=Prot;produc
NC_045512.2 RefSeq  mature_protein_region_of_CDS    806 2719    .   +   .   ID=id-YP_009724389.1:181..818;Note=produced by both ppla and pplab;Parent=cds-YP_009724389.1;gbkey=Prot;product=
NC_045512.2 RefSeq  mature_protein_region_of_CDS    2720    8554    .   +   .   ID=id-YP_009724389.1:819..2763;Note=former nsp1%3B conserved domains are: N-terminal acidic (Ac)%2C predicte
NC_045512.2 RefSeq  mature_protein_region_of_CDS    8555    10054   .   +   .   ID=id-YP_009724389.1:2764..3263;Note=nsp4B_TM%3B contains transmembrane domain 2 (TM2)%3B produced by both p
NC_045512.2 RefSeq  mature_protein_region_of_CDS    10055   10972   .   +   .   ID=id-YP_009724389.1:3264..3569;Note=nsp5A_3CLpro and nsp5B_3CLpro%3B main proteinase (Mpro)%3B mediates cle
NC_045512.2 RefSeq  mature_protein_region_of_CDS    10973   11842   .   +   .   ID=id-YP_009724389.1:3570..3859;Note=nsp6_TM%3B putative transmembrane domain%3B produced by both ppla and p
NC_045512.2 RefSeq  mature_protein_region_of_CDS    11843   12091   .   +   .   ID=id-YP_009724389.1:3860..3942;Note=produced by both ppla and pplab;Parent=cds-YP_009724389.1;gbkey=Prot;pr
NC_045512.2 RefSeq  mature_protein_region_of_CDS    12092   12685   .   +   .   ID=id-YP_009724389.1:3943..4140;Note=produced by both ppla and pplab;Parent=cds-YP_009724389.1;gbkey=Prot;pr
NC_045512.2 RefSeq  mature_protein_region_of_CDS    12686   13024   .   +   .   ID=id-YP_009724389.1:4141..4253;Note=ssRNA-binding protein%3B produced by both ppla and pplab;Parent=cds-YP_
NC_045512.2 RefSeq  mature_protein_region_of_CDS    13025   13441   .   +   .   ID=id-YP_009724389.1:4254..4392;Note=nsp10_CysHis%3B formerly known as growth-factor-like protein (GFL)%3B p
NC_045512.2 RefSeq  mature_protein_region_of_CDS    13442   13468   .   +   .   ID=id-YP_009724389.1:4393..5324;Note=nsp12%3B NiRAN and RdRp%3B produced by pplab only;Parent=cds-YP_0097243
NC_045512.2 RefSeq  mature_protein_region_of_CDS    13468   16236   .   +   .   ID=id-YP_009724389.1:4393..5324;Note=nsp12%3B NiRAN and RdRp%3B produced by pplab only;Parent=cds-YP_0097243
NC_045512.2 RefSeq  mature_protein_region_of_CDS    16237   18039   .   +   .   ID=id-YP_009724389.1:5325..5925;Note=nsp13_ZBD%2C nsp13_TB%2C and nsp_HEL1core%3B zinc-binding domain (ZD)%2
NC_045512.2 RefSeq  mature_protein_region_of_CDS    18040   19620   .   +   .   ID=id-YP_009724389.1:5926..6452;Note=nsp14A2_ExoN and nsp14B_NMT%3B produced by pplab only;Parent=cds-YP_009
NC_045512.2 RefSeq  mature_protein_region_of_CDS    19621   20658   .   +   .   ID=id-YP_009724389.1:6453..6798;Note=nsp15-A1 and nsp15B-NendoU%3B produced by pplab only;Parent=cds-YP_0097
NC_045512.2 RefSeq  mature_protein_region_of_CDS    20659   21552   .   +   .   ID=id-YP_009724389.1:6799..7096;Note=nsp16_OMT%3B 2'-o-MT%3B produced by pplab only;Parent=cds-YP_009724389.
NC_045512.2 RefSeq  CDS 266 13483   .   +   0   ID=cds-YP_009725295.1;Parent=gene-GU280_gp01;Dbxref=Genbank:YP_009725295.1,GeneID:43740578;Name=YP_009725295.1;Note=ppla;gbkey=CDS;gene=ORF1
NC_045512.2 RefSeq  mature_protein_region_of_CDS    266 805 .   +   .   ID=id-YP_009725295.1:1..180;Note=nsp1%3B produced by both ppla and pplab;Parent=cds-YP_009725295.1;gbkey=Prot;produc
NC_045512.2 RefSeq  mature_protein_region_of_CDS    806 2719    .   +   .   ID=id-YP_009725295.1:181..818;Note=produced by both ppla and pplab;Parent=cds-YP_009725295.1;gbkey=Prot;product=
NC_045512.2 RefSeq  mature_protein_region_of_CDS    2720    8554    .   +   .   ID=id-YP_009725295.1:819..2763;Note=former nsp1%3B conserved domains are: N-terminal acidic (Ac)%2C predicte
NC_045512.2 RefSeq  mature_protein_region_of_CDS    8555    10054   .   +   .   ID=id-YP_009725295.1:2764..3263;Note=nsp4B_TM%3B contains transmembrane domain 2 (TM2)%3B produced by both p
NC_045512.2 RefSeq  mature_protein_region_of_CDS    10055   10972   .   +   .   ID=id-YP_009725295.1:3264..3569;Note=nsp5A_3CLpro and nsp5B_3CLpro%3B main proteinase (Mpro)%3B mediates cle
NC_045512.2 RefSeq  mature_protein_region_of_CDS    10973   11842   .   +   .   ID=id-YP_009725295.1:3570..3859;Note=nsp6_TM%3B putative transmembrane domain%3B produced by both ppla and p
NC_045512.2 RefSeq  mature_protein_region_of_CDS    11843   12091   .   +   .   ID=id-YP_009725295.1:3860..3942;Note=produced by both ppla and pplab;Parent=cds-YP_009725295.1;gbkey=Prot;pr
NC_045512.2 RefSeq  mature_protein_region_of_CDS    12092   12685   .   +   .   ID=id-YP_009725295.1:3943..4140;Note=produced by both ppla and pplab;Parent=cds-YP_009725295.1;gbkey=Prot;pr
NC_045512.2 RefSeq  mature_protein_region_of_CDS    12686   13024   .   +   .   ID=id-YP_009725295.1:4141..4253;Note=ssRNA-binding protein%3B produced by both ppla and pplab;Parent=cds-YP_
NC_045512.2 RefSeq  mature_protein_region_of_CDS    13025   13441   .   +   .   ID=id-YP_009725295.1:4254..4392;Note=nsp10_CysHis%3B formerly known as growth-factor-like protein (GFL)%3B p
NC_045512.2 RefSeq  mature_protein_region_of_CDS    13442   13480   .   +   .   ID=id-YP_009725295.1:4393..4405;Note=produced by ppla only;Parent=cds-YP_009725295.1;gbkey=Prot;product=nspl
NC_045512.2 RefSeq  stem_loop   13476   13503   .   +   .   ID=id-GU280_gp01;Dbxref=GeneID:43740578;function=Coronavirus frameshifting stimulation element stem-loop 1;gbkey=stem_loop;gene=
NC_045512.2 RefSeq  stem_loop   13488   13542   .   +   .   ID=id-GU280_gp01-2;Dbxref=GeneID:43740578;function=Coronavirus frameshifting stimulation element stem-loop 2;gbkey=stem_loop;ger
NC_045512.2 RefSeq  gene    21563   25384   .   +   .   ID=gene-GU280_gp02;Dbxref=GeneID:43740568;Name=S;gbkey=Gene;gene=S;gene_biotype=protein_coding;gene_synonym=spike glycoprotein;locus
NC_045512.2 RefSeq  CDS 21563   25384   .   +   0   ID=cds-YP_009724390.1;Parent=gene-GU280_gp02;Dbxref=Genbank:YP_009724390.1,GeneID:43740568;Name=YP_009724390.1;Note=structural protein%3
NC_045512.2 RefSeq  gene    25393   26220   .   +   .   ID=gene-GU280_gp03;Dbxref=GeneID:43740569;Name=ORF3a;gbkey=Gene;gene=ORF3a;gene_biotype=protein_coding;locus_tag=GU280_gp03
NC_045512.2 RefSeq  CDS 25393   26220   .   +   0   ID=cds-YP_009724391.1;Parent=gene-GU280_gp03;Dbxref=Genbank:YP_009724391.1,GeneID:43740569;Name=YP_009724391.1;gbkey=CDS;gene=ORF3a;locu
NC_045512.2 RefSeq  gene    26245   26472   .   +   .   ID=gene-GU280_gp04;Dbxref=GeneID:43740570;Name=E;gbkey=Gene;gene=E;gene_biotype=protein_coding;locus_tag=GU280_gp04
NC_045512.2 RefSeq  CDS 26245   26472   .   +   0   ID=cds-YP_009724392.1;Parent=gene-GU280_gp04;Dbxref=Genbank:YP_009724392.1,GeneID:43740570;Name=YP_009724392.1;Note=ORF4%3B structural p
NC_045512.2 RefSeq  gene    26523   27191   .   +   .   ID=gene-GU280_gp05;Dbxref=GeneID:43740571;Name=M;gbkey=Gene;gene=M;gene_biotype=protein_coding;locus_tag=GU280_gp05
NC_045512.2 RefSeq  CDS 26523   27191   .   +   0   ID=cds-YP_009724393.1;Parent=gene-GU280_gp05;Dbxref=Genbank:YP_009724393.1,GeneID:43740571;Name=YP_009724393.1;Note=ORF5%3B structural p
NC_045512.2 RefSeq  gene    27202   27387   .   +   .   ID=gene-GU280_gp06;Dbxref=GeneID:43740572;Name=ORF6;gbkey=Gene;gene=ORF6;gene_biotype=protein_coding;locus_tag=GU280_gp06
NC_045512.2 RefSeq  CDS 27202   27387   .   +   0   ID=cds-YP_009724394.1;Parent=gene-GU280_gp06;Dbxref=Genbank:YP_009724394.1,GeneID:43740572;Name=YP_009724394.1;gbkey=CDS;gene=ORF6;locus
NC_045512.2 RefSeq  gene    27394   27759   .   +   .   ID=gene-GU280_gp07;Dbxref=GeneID:43740573;Name=ORF7a;gbkey=Gene;gene=ORF7a;gene_biotype=protein_coding;locus_tag=GU280_gp07
NC_045512.2 RefSeq  CDS 27394   27759   .   +   0   ID=cds-YP_009724395.1;Parent=gene-GU280_gp07;Dbxref=Genbank:YP_009724395.1,GeneID:43740573;Name=YP_009724395.1;gbkey=CDS;gene=ORF7a;locu
NC_045512.2 RefSeq  gene    27756   27887   .   +   .   ID=gene-GU280_gp08;Dbxref=GeneID:43740574;Name=ORF7b;gbkey=Gene;gene=ORF7b;gene_biotype=protein_coding;locus_tag=GU280_gp08
NC_045512.2 RefSeq  CDS 27756   27887   .   +   0   ID=cds-YP_009725318.1;Parent=gene-GU280_gp08;Dbxref=Genbank:YP_009725318.1,GeneID:43740574;Name=YP_009725318.1;gbkey=CDS;gene=ORF7b;locu
NC_045512.2 RefSeq  gene    27894   28259   .   +   .   ID=gene-GU280_gp09;Dbxref=GeneID:43740577;Name=ORF8;gbkey=Gene;gene=ORF8;gene_biotype=protein_coding;locus_tag=GU280_gp09
NC_045512.2 RefSeq  CDS 27894   28259   .   +   0   ID=cds-YP_009724396.1;Parent=gene-GU280_gp09;Dbxref=Genbank:YP_009724396.1,GeneID:43740577;Name=YP_009724396.1;gbkey=CDS;gene=ORF8;locus
NC_045512.2 RefSeq  gene    28274   29533   .   +   .   ID=gene-GU280_gp10;Dbxref=GeneID:43740575;Name=N;gbkey=Gene;gene=N;gene_biotype=protein_coding;locus_tag=GU280_gp10
```